An IEEE Societies and Technical Councils Publication
IEEE Data Descriptions

Received 12 September 2024; revised 24 September 2024, 31 October 2024, and 3 November 2024; accepted 4 November 2024;
Date of publication 8 November, 2024; date of current version 11 December, 2024.

Digital Object Identifier 10.1109/IEEEDATA.2024.3493798

Descriptor: *Voice Pre-Processing and Quality Assessment Dataset (VPQAD)*

AJAN AHMED (GRADUATE STUDENT MEMBER, IEEE), MD JAHANGIR ALAM KHONDKAR, ANSEN HERRICK, STEPHANIE SCHUCKERS (FELLOW, IEEE), AND MASUDUL H. IMTIAZ (MEMBER, IEEE)

Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY 13699 USA

CORRESPONDING AUTHOR: Masudul H. Imtiaz (e-mail: mimtiaz@clarkson.edu).

This work was supported by the Center for Identification Technology Research and the National Science Foundation under Grant 1650503.

ABSTRACT This article introduces the voice pre-processing and quality assessment dataset (VPQAD), a scalable resource developed to validate various pre-processing techniques and improve voice signal quality in noisy environments. The dataset comprises voice recordings from 50 participants aged 18–40, captured in controlled real-life conditions using Audio Technica AT2020 and SHURE SM58 microphones. These high-quality recordings, made under diverse noise levels and settings, could be used for testing and developing voice enhancement algorithms. The dataset includes detailed metadata on the environment and participant demographics for analyzing and improving speech clarity and intelligibility, particularly in challenging conditions. To protect privacy, all data have been anonymized. VPQAD has been made public to promote collaborative research and advance research in biometrics, telecommunications, assistive technologies, and other applications requiring clear voice communication.

IEEE SOCIETY/COUNCIL Signal Processing Society (SPS)

DATA TYPE/LOCATION Audio; Potsdam, NY, USA

DATA DOI/PID 10.21227/yb1h-hs38

INDEX TERMS Automated speaker recognition (ASR), noise-robust ASR systems, speech quality assessment, text-dependent speech, text-independent speech, voice datasets, voice pre-processing.

BACKGROUND

Developing reliable voice processing technologies has become increasingly crucial in our digitally connected world, where clear and intelligible speech communication is essential [1]. Background noise and poor recording environments may degrade the quality of voice signals, impacting various applications from telecommunications to voice-controlled systems [2]. For instance, speaker recognition systems may struggle to differentiate between speakers when there is a high level of ambient noise, which can reduce the accuracy of the recognition process [3], [4].

Over the years, several datasets have been developed to support research in voice processing. However, many resources are limited, often focusing on controlled environments or lacking the diversity needed to simulate real-world noise conditions [5]. These datasets have been invaluable in advancing research. Still, their inability to capture the full complexity of modern-day voice applications has created a need for more robust and comprehensive datasets [6].

In response to this need, we introduce the voice pre-processing and quality assessment dataset (VPQAD), a comprehensive resource developed to facilitate research in improving voice signal quality in noisy environments. VPQAD comprises voice recordings from 50 participants captured using high-quality microphones in controlled, real-world noise conditions. Additionally, VPQAD includes both text-dependent and text-independent speech from each participant, allowing for various analyses and applications.

While VPQAD serves as a general-purpose dataset for enhancing voice signal clarity, its most significant application could be advancing automated speaker recognition (ASR) systems. ASR technology, vital for security, authentication, and numerous other applications, depends heavily on the quality of input signals [7]. Noise and poor recording conditions can significantly impair ASR accuracy, leading to incorrect speaker identification and reduced system reliability [8].

In this article, we discuss the data collection process, the variety of noise environments, and the metadata accompanying each recording. Additionally, we explore how VPQAD can be used to push the boundaries of current ASR technology, enabling more accurate speaker recognition in real-world scenarios.

Related Work

One of the earliest and most widely used datasets is the *TIMIT* Acoustic-Phonetic Continuous Speech Corpus, developed in the late 1980s. TIMIT is renowned for its phonetic richness, making it ideal for phoneme recognition and speech synthesis. However, its clean recording conditions limit its effectiveness in developing noise-robust ASR systems [9].

LibriSpeech, a large-scale corpus derived from public domain audiobooks, offers over 1000 hours of read English speech. It is widely used as a standard ASR benchmark. Despite its extensive coverage, LibriSpeech primarily features read speech and lacks the environmental noise diversity necessary for training robust ASR systems [10].

The Aurora project introduced a series of datasets such as *Aurora-2* and *Aurora-4*, specifically designed to test the noise robustness of ASR systems. These datasets are valuable for developing noise-robust algorithms because they focus on degraded speech quality. However, they are limited by the artificial nature of the introduced noise and a lack of linguistic diversity [11].

The *AMI Meeting Corpus* provides a rich resource for research involving multi-party meetings in various acoustic environments. This dataset, which includes audio, video, and transcripts, is particularly relevant for ASR and speaker identification in noisy, multi-speaker settings. However, its complexity requires sophisticated models to achieve high accuracy [12].

Mozilla's Common Voice dataset stands out for its linguistic diversity, with recordings in over 70 languages, making it one of the most inclusive datasets available. However, the variability in recording quality due to the wide range of devices and environments used by contributors presents challenges for ASR development [13].

VoxCeleb is a large-scale speaker identification dataset with speech samples extracted from YouTube videos. It captures various acoustic environments, ranging from studio-quality interviews to noisy public events. While VoxCeleb is particularly useful for speaker recognition and verification tasks, its focus on speaker identification limits its broader applicability to general ASR research [14].

Finally, the *CHiME* challenge series provides datasets designed to test ASR systems in noisy, everyday environments such as public transport and cafes. The CHiME datasets, while valuable for their intended applications, focus on multi-microphone, far-field conditions, making them less suitable for tasks involving single-microphone recordings or specific noise environments such as low-bandwidth audio [15].

Table I compares VPQAD with existing datasets, highlighting their contributions and limitations in ASR and speech processing. While each dataset has advanced the field by addressing challenges such as noisy environments and diverse speaker populations, gaps remain, particularly in real-world noise conditions and linguistic diversity.

Contributions of This Article

One of VPQAD's most significant contributions is its dual approach to capturing text-dependent speech (the same specific phrases or words are used by speakers [3]) and text-independent speech (speakers can use any words or phrases [16]) from each participant to allow diverse ASR research and development needs. Text-dependent speech is crucial for security applications relying on specific authentication phrases, such as voice-based systems, where users must say a predefined password such as "apple" for authentication. In contrast, text-independent speech supports broader applications, such as speaker verification in customer service call centers, where the system must recognize speakers regardless of their words.

Another significant contribution is the incorporation of real-world noise into the recordings. Unlike many datasets that add synthetic noise to otherwise clean recordings, VPQAD's audio data are captured in naturally noisy environments, such as cafeterias and laboratories during active sessions. This aspect is critical for advancing noise-robust ASR systems, as it better simulates the actual conditions in which they are expected to operate.

VPQAD also emphasizes balanced speaker representation, featuring recordings from 50 diverse participants. This balance is essential for training ASR models that generalize well across different voices, preventing overfitting to specific demographic characteristics and thereby reducing bias. In addition to diversity in speakers, VPQAD includes a mix of speech styles—from conversational to naming objects.

Furthermore, this article introduces proprietary software specifically developed to facilitate the data collection process for VPQAD. This software automates the presentation of prompts and images to participants, ensuring consistency in data collection across sessions. By providing a controlled and repeatable environment for eliciting text-dependent and text-independent speech, the software plays a crucial role in the quality and reliability of the VPQAD dataset.

TABLE I. Summary of the VPQAD Dataset Versus Prior Work

Dataset	Public	Audio Type	Background Noise	Number of Speakers	Language	Speech Style	Recording Environment
VPQAD	**Yes**	**Text-Dependent and Independent**	**Real-World Scenarios**	**50**	**English**	**Mixed**	**Cafeterias and Laboratories during lab classes**
TIMIT	No	Text-Dependent	None	630	English	Read Speech	Studio
LibriSpeech	Yes	Text-Dependent	None	2456	English	Read Speech	Various
Aurora-4	No	Text-Dependent	Artificially Added	83	English	Read Speech	Synthetic and Real-World
AMI Meeting Corpus	No	Text-Independent	Real-World Scenarios	100+	English	Conversational	Office and Meeting Rooms
Common Voice	Yes	Text-Dependent	Real-World Scenarios	70 000+	Multilingual	Read Speech	Various
VoxCeleb	No	Text-Independent	Real-World Scenarios	7000+	English	Conversational	Various
CHiME-3	No	Text-Independent	Real-World Scenarios	Various	English	Conversational	Public Transport and Cafes

Note: This table focuses on research use, speaker diversity, language, speech style, recording environment, and the inclusion of background noise, categorized as artificially added or from real-world scenarios.

The dataset's design also supports future expansion, particularly in the multilingual domain. While VPQAD currently focuses on English, its framework and methodology provide a solid foundation for including other languages. This potential for expansion is crucial as the demand for multilingual ASR systems grows globally, making VPQAD a potentially invaluable resource for researchers working on ASR technologies that need to accommodate multiple languages.

Additionally, GitHub links to Matlab and Python scripts created by the authors to measure audio quality and edit audio clips are provided.

COLLECTION METHODS AND DESIGN

Institutional Review Board (IRB) Approval

VPQAD was developed following rigorous ethical guidelines and procedures approved by the Institutional Review Board (IRB Approval No. 24-42) at Clarkson University [17]. This approval ensures that all aspects of the research involving human subjects adhere to the highest ethical standards, particularly regarding informed consent, data confidentiality, and the overall treatment of participants.

Ethical Considerations and Informed Consent

The IRB's primary role is to safeguard the rights and well being of research participants. All participants were fully informed about the study's nature, potential risks and benefits, and their rights as participants. Each participant signed an informed consent form detailing the study's purpose, procedures, data-sharing permissions, and measures to ensure their confidentiality. Flyers were distributed throughout the university campus to recruit participants.

Data Confidentiality and Security

The IRB addressed the critical concern of protecting participant data. All voice recordings and associated metadata were anonymized before inclusion in the dataset, meaning no identifying information was linked to the tapes. After anonymization, all identifiable data were permanently deleted, and the consent forms were physically and securely stored at the university. The data were stored in secure, password-protected environments, accessible only to authorized researchers.

Participant Recruitment and Consent

Participants for the VPQAD dataset were recruited from the Clarkson University community through flyers and electronic communications. All potential participants were provided with detailed information about the study, including its objectives, procedures, and their rights as participants. Informed consent was obtained from all participants before the data collection began.

Recording Environments

Recordings were conducted in controlled, naturally noisy environments. The primary locations for data collection include the following.

University Cafeteria

Data were collected at lunchtime between 12 and 2 pm when the cafeteria is most filled during the day. Background noises include conversations, the clinking of utensils and plates, footsteps, background music, kitchen sounds, cash registers, payment beeps, moving chairs, ambient noise, and outside noise.

Laboratories During Lab Sessions

Data were collected during lab sessions between 12 and 2 pm. The lab classes had about 25–30 students conducting their lab sessions. Background noises include conversations between students and instructors, the buzzing or beeping of electronic devices, the hum of power supplies or transformers, the clicking of switches or buttons, the whirring of cooling vents, the clinking of components being handled, the sound of keyboards and mice, etc.

Fig. 1 shows examples of these environments' live data collection process.

FIG. 1. **Live data collection in electrical engineering lab (top) and University Cafeteria (bottom).**

Recording Procedure Using Custom Software

The data collection software was built using Python, using the Pyaudio library for real-time audio capture, at a sampling rate of 44 100 samples per second. The software includes the functionality to automatically detect available microphones on the system and allows the user to select one or two microphones depending on their recording needs. If no selection is made, a default microphone is activated.

The software's graphical user interface (GUI) was designed to display images and instructional text during the recording process. The GUI integrates features such as microphone status indicators, customizable image sequences, and a settings menu for configuring recording parameters. The threading functionality in Python enables the simultaneous recording of audio from multiple microphones. Additionally, the software incorporates logic for window resizing, allowing it to be used across various screen sizes and resolutions without compromising the clarity of displayed text or images.

The contributed software presented images to be named out loud and text prompts on the screen. This process was divided into two key phases as follows.

Text-Dependent Phase

Participants were shown a series of images (e.g., a flower, a frog, and a bike) along with corresponding text labels and were instructed to say the name of each object out loud. Each audio clip consists of 30 s of data collected for every participant in each session. Ten spoken words were within that time frame: flower, frog, bike, car, fork, cinnamon, pizza, broccoli, chair, and bear. The following script can be used to trim this clip into smaller segments containing individual words spoken: https://github.com/ahmedajan/SNR_Calculation_For_VPQAD/blob/main/audio_Segmentation.m.

Text-Independent Phase

Participants were also given prompts that required spontaneous speech, such as describing a course they liked recently or their research. A total of 20 s of data were collected for every participant in each session.

This 20-s window reflects real-world constraints, ensuring that models can perform effectively with limited data. Additionally, collecting 20 s of speech per participant is often sufficient because it captures essential voice characteristics, such as pitch, timbre, and speaking rate, which are crucial for speaker recognition [18].

The example of images that showed up on the interface of the custom software used during these sessions is shown in Fig. 2.

Recording Equipment

All voice recordings were captured using the Audio Technica AT2020 [19] and SHURE SM58 [20] models shown

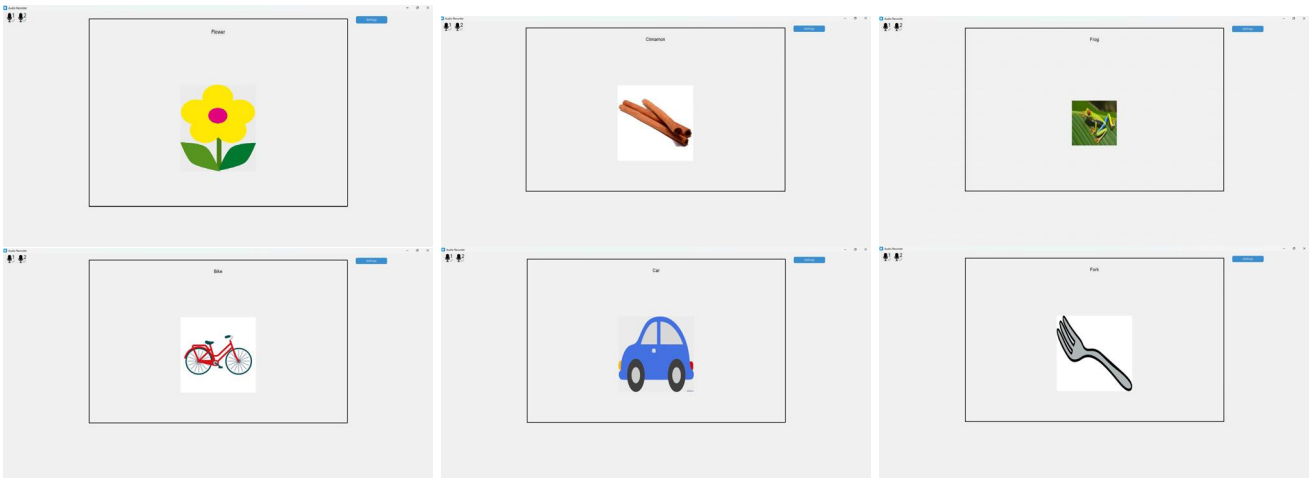

FIG. 2. **Screenshots of the images that showed up on the interface of the software during data collection.**

FIG. 3. **Microphones used in the recording: Audio Technica AT2020 (left) and SHURE SM58 (right).**

in Fig. 3. Their technical specifications can be found on each of the manufacturer's respective commercial websites [19], [20].

The recordings were made at a standard sampling rate of 44.1 kHz. The Audio-Technica AT2020 operates within a frequency range of 20 Hz–20 kHz, has a sensitivity of −37 dB, and can handle a maximum sound pressure level (SPL) of 144 dB. It requires 48 V phantom power to operate and connects via an XLR output [19]. The Shure SM58 operates within a frequency response range of 50 Hz–15 kHz and is optimized to emphasize clarity in the vocal midrange while attenuating low-frequency background noise [20].

Dataset Structure

VPQAD is organized into directories based on recording sessions, with each session contains the following.

1) *Text-Dependent Recordings*: Stored in `td/`.
2) *Text-Independent Recordings*: Stored in `tid/`.

Speech Content

VPQAD includes the following.

1) *Text-Dependent*: Recordings with specific objects being named out loud. Fig. 4 shows a waveform of text-dependent data where an energy-based recognition script detects spoken words.

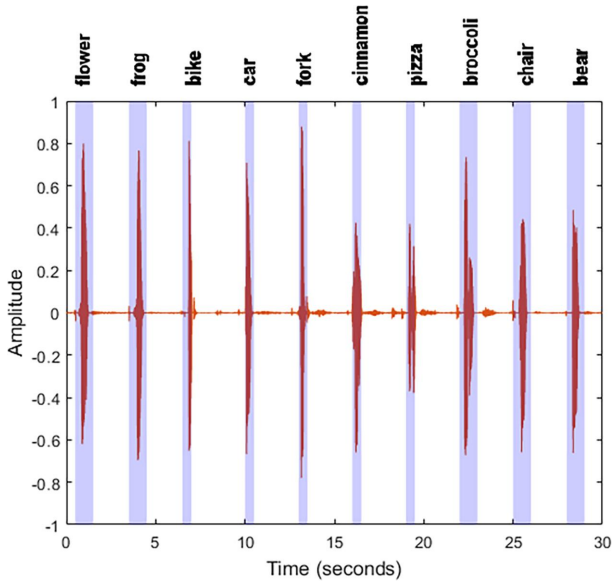

FIG. 4. **Waveform of a text-dependent data. The purple-shaded frames are the ones where speech is detected.**

2) *Text-Independent Recordings*: Free-form speech of people uniquely describing their recent course or research.

VALIDATION AND QUALITY

Data Quality Control

Each recording was manually reviewed for clarity and consistency to ensure the highest quality data. The final dataset excluded records that contained any clipping, inconsistent sampling rates (44.1 kHz), unclear or interrupted speech, and technical malfunctions.

Speaker Diversity

As shown in Table II, the dataset includes recordings from 50 participants of diverse backgrounds, ensuring broad

TABLE II. Participant Demographic Information

Category	Subcategory	Number of Participants
Gender	Man	39
	Woman	9
	Prefer not to respond	2
Race	Caucasian	31
	Black	8
	Hispanic	3
	Native American	1
	Middle Eastern	1
	Indian	3
	Asian	2
	Other	1
Age	18–25 years	39
	26–30 years	6
	31–40 years	5

TABLE III. Summary of Audio Properties in the VPQAD Dataset

Metric	Value
Sampling Rate	44.1 kHz
Bit Depth	16-bit
Average Noise Level of All Audio Files	19.60 dB
Total Duration of Recordings	68 min

representation across age, gender, and accents, which is crucial for training ASR models that generalize well.

Audio Quality Metrics

The audio properties highlighted in Table III demonstrate the key attributes. Each participant recorded a total of 50 s, including both text-dependent and text-independent data in each session. Noise levels were measured in A-weighted decibels (dBAs) using calibrated microphones. The noise levels ranged between 4.86 and 62 dBA, with an average of 19.60 dBA. This range reflects noise from near-silent environments (4.86 dB) to moderately noisy settings such as normal conversation (62 dB) [21], [22], [23].

Table IV shows various SNR values calculated. The signal-to-noise ratio (SNR) was calculated using the following formula:

$$\text{SNR (in dB)} = 10 \times \log_{10}\left(\frac{P_{\text{signal}}}{P_{\text{noise}}}\right)$$

where P_{signal} is the power of the speech signal and P_{noise} is the power of the background noise. The power for both the signal and noise was computed as the mean square of their respective amplitudes

$$P = \frac{1}{N}\sum_{n=1}^{N} x[n]^2$$

where $x[n]$ represents the amplitude of the signal (for P_{signal}) or the noise (for P_{noise}) at sample n, and N is the total number of samples.

TABLE IV. Summary of Different SNR for VPQAD

Metric	Highest Value (dB)	Lowest Value (dB)	Mean Value (dB)
SNR	45.22	15.35	22.57
Segmented SNR (SegSNR)	42.14	18.73	24.92
Frequency-Weighted SNR	44.50	15.60	23.86

To estimate noise, the code assumes that the low-energy portions of the signal correspond to noise. Specifically, any segment of the audio where the amplitude falls below a predefined threshold (based on the mean signal energy) is considered noise, and the rest is considered the speech signal. The signal-to-noise ratio was also calculated using two additional methods: segmented SNR (SegSNR) and frequency-weighted SNR (fwSNR). For SegSNR, the signal-to-noise ratio is computed over short, fixed-length segments of the signal. Each segment's SNR is calculated using the formula

$$\text{SegSNR (in dB)} = 10 \times \log_{10}\left(\frac{P_{\text{segment}}}{P_{\text{noise}}}\right)$$

where P_{segment} is the power of the speech signal in each segment, and P_{noise} is the power of the noise. The final SegSNR is the average of the SNR values over all segments. The segment lengths are chosen based on typical speech durations, often around 20–30 ms, which are sufficient to capture phoneme-level details in speech. In our code, noise is estimated by identifying low-energy segments of the signal, assuming that these segments correspond to noise-dominated regions, based on a threshold relative to the overall signal energy.

For fwSNR, the A-weighting filter is used to emphasize frequency bands that are more important for human hearing. This filter applies greater weight to mid-frequencies (500 Hz–5 kHz) and attenuates lower and higher frequencies. The formula is as follows:

$$\text{fwSNR (in dB)} = 10 \times \log_{10}\left(\frac{P_{\text{A-weighted signal}}}{P_{\text{A-weighted noise}}}\right)$$

where $P_{\text{A-weighted signal}}$ and $P_{\text{A-weighted noise}}$ are the power values of the signal and noise after applying the A-weighting filter. In the code, noise is estimated similarly to SegSNR, by identifying low-energy portions of the audio that fall below a set threshold. The scripts used for SNR calculations can be found here: https://github.com/ahmedajan/SNR_Calculation_For_VPQAD/tree/main.

RECORDS AND STORAGE

Data Processing and Storage

After each recording session, the data were stored in a secure, password-protected digital archive, with access only to authorized researchers. This was done to maintain privacy

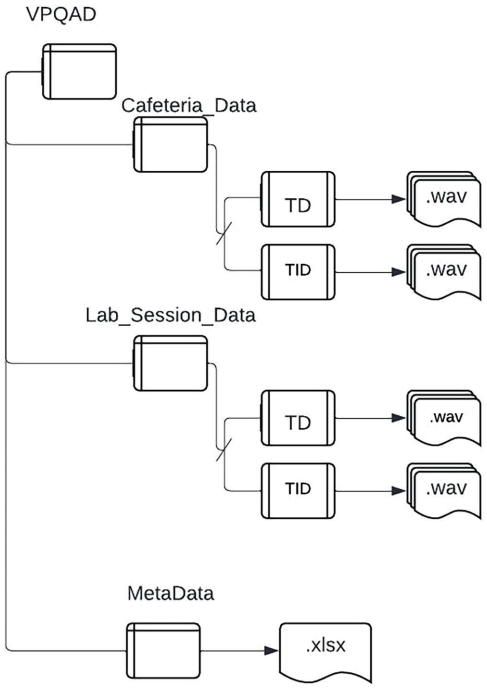

FIG. 5. Data directory architecture. TD represents the folder with text-dependent data, and TID represents the folder with text-independent data.

while all identifying information was separated from the audio files and permanently deleted from the recordings. The files were anonymized by assigning each recording a unique identification code, ensuring no personal information was linked to the voice data. This ensured that the dataset could now be made public while maintaining participants' privacy according to the IRB regulations.

File Naming Conventions and Folder Architecture

Files are named subjectID_ microphoneNumber_ RecordingType.wav (e.g., sub001_1_td.wav). Fig. 5 shows the Directory architecture. Microphone number 1 is the AT2020, and microphone number 2 represents the SM58. For recording type, td represents text-dependent data, and tid represents text-independent data.

INSIGHTS AND NOTES

Accessing Data

The VPQAD is intended exclusively for academic research. To obtain access, researchers are required to sign the End User License Agreement (EULA), which can be requested via email at mimtiaz@clarkson.edu or downloaded directly from the IEEE Dataport. A signed EULA must then be returned to this email address. Only emails originating from academic accounts will be accepted.

Dataset Limitations

With VPQAD, several limitations should be considered. The mean noise level of 19.60 dB suggests that most of the data were collected in moderately noisy environments, similar to what might be experienced in university settings. While this range is suitable for simulating low to moderate noise conditions, it does not fully capture highly noisy environments such as urban streets and industrial settings, which typically feature noise above 70 dB. Therefore, while adequate for many real-world scenarios, future data collection efforts may consider incorporating higher noise levels to test system performance in more challenging acoustic conditions.

The 44.1 kHz sampling rate and 16-bit bit depth are commonly used in audio data collection, capturing the full range of human speech and providing adequate dynamic range. However, to improve the precision of data, increasing the sampling rate to 48 kHz could allow for a better representation of high-frequency noise, which may be present in more complex acoustic environments [21], [24]. Additionally, upgrading to 24-bit depth would enhance the dynamic range, allowing for more accurate handling of lower level signals in noisy environments and thus improving speech recognition and enhancement models [22].

The dataset focuses exclusively on English, which may limit its use for developing multilingual ASR systems. Additionally, although it includes recordings in real-world noisy environments, these conditions are still somewhat controlled and may not fully capture the complexity of all real-world scenarios. The participant pool, though diverse, is limited to 50 individuals, which may not represent broader population demographics. Moreover, using consistent, high-quality recording equipment, while beneficial for audio clarity, could introduce a bias toward specific acoustic characteristics, potentially affecting the generalizability of ASR models. Finally, while the dataset's total duration of 50 h is substantial, it may still be insufficient for training large-scale ASR models, necessitating additional data sources.

SOURCE CODE AND SCRIPTS

The scripts for calculating signal-to-noise ratios (SNRs), segmenting audio files, and evaluating speech quality metrics are publicly accessible. These also include segmentation scripts that can trim recordings into smaller segments containing individual spoken words. These are available via GitHub and the repository link for scripts used in this dataset is available at: https://github.com/ahmedajan/SNR_Calculation_For_VPQAD/tree/main.

ACKNOWLEDGMENTS

The authors give special thanks to Zahra Mahdavi for her help in data collection.

A.A. and M.J.A.K. collected, curated, and analyzed the data. A.A. wrote the manuscript. A.H. developed the software used in data collection. S.S. and M.H.I. reviewed the data collection, curation, and analysis. All authors reviewed the manuscript.

The authors have declared no conflicts of interest.

REFERENCES

[1] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2018, pp. 5024–5028.

[2] J. L. K. Fendji, D. M. Tala, B. O. Yenke, and M. Atemkeng, "Automatic speech recognition using limited vocabulary: A survey," 2021, *arXiv:2108.10254*.

[3] T. Kinnunen and H. Li, "An overview of speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio," in *New Era for Robust Speech Recognition*. Cham, Switzerland: Springer, 2017, pp. 165–186.

[5] J. R. Hershey, S. J. Rennie, M. Aharon, and R. Gopinath, "Superhuman multi-talker speech recognition: A graphical modeling approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Piscataway, NJ, USA: IEEE Press, 2010, pp. 4398–4401.

[6] B. Elizalde, C. Zhang, V. Jayaram, L. Sigal, and B. Raj, "Cross-dataset generalization in automatic speech recognition: From controlled lab to real-world data," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 6699–6703.

[7] M. Dua, C. Jain, and S. Kumar, "LSTM and CNN based ensemble approach for spoof detection task in automatic speaker verification systems," *J. Ambient Intell. Humanized Comput.*, vol. 13, pp. 1–16, Apr. 2022.

[8] H. Dubey, A. Sangwan, and J. H. Hansen, "Leveraging frequency-dependent kernel and dip-based clustering for robust speech activity detection in naturalistic audio streams," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2056–2071, Nov. 2018.

[9] J. S. Garofolo et al., "TIMIT: Acoustic-phonetic continuous speech corpus," WorldCat, 1993. Accessed: Sep. 18, 2024. [Online]. Available: http://www.worldcat.org/isbn/1585630195

[10] D. P. V. Panayotov, G. Chen and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 5206–5210.

[11] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000, pp. 29–32.

[12] J. Carletta et al., "The AMI meeting corpus," in *Proc. Int. Conf. Methods Techn. Behav. Res.*, 2005.

[13] "Mozilla common voice dataset," Mozilla Common Voice, 2023. Accessed: Sep. 24, 2024. [Online]. Available: https://commonvoice.mozilla.org/en/datasets

[14] J. S. C. A. Nagrani and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[15] E. V. J. Barker, R. Marxer and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task, and baselines," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2015, pp. 126–130.

[16] J. P. Campbell, "Speaker recognition: A tutorial," in *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[17] "Institutional Review Board (IRB)," Clarkson University, 2024. Accessed: Aug. 8, 2024. [Online]. Available: https://www.clarkson.edu/academics/research/institutional-review-board

[18] J. H. Hansen and T. Hasan, "Speaker recognition: A tutorial," in *Springer Handbook of Speech Processing*. Berlin, Germany: Springer, 2015, pp. 305–329.

[19] "At2020 cardioid condenser microphone," Audio-Technica, 2024. Accessed: Oct. 24, 2024. [Online]. Available: https://www.audio-technica.com/en-us/at2020

[20] "Sm58 vocal microphone," Shure, 2024. Accessed: Aug. 20, 2024. [Online]. Available: https://www.shure.com/en-US/products/microphones/sm58?variant=SM58-LC

[21] E. McPhillips, "Noise levels of everyday sounds," Audicus, 2022. Accessed: Sep. 5, 2024. [Online]. Available: https://www.audicus.com/noise-levels-of-everyday-sounds/

[22] "Decibel examples: Noise levels of common sounds," Lexie Hearing Aids, 2024. Accessed: Sep. 5, 2024. [Online]. Available: https://lexiehearing.com/blog/decibel-examples

[23] "Common noise levels," Noiseawareness.org, 2024. Accessed: Sep. 5, 2024. [Online]. Available: https://noiseawareness.org/info-center/common-noise-levels/

[24] "Understanding decibel levels for hearing health," Soundly, 2024. Accessed: Sep. 5, 2024. [Online]. Available: https://www.soundly.com/noise-levels